

(19) 日本国特許庁 (J P)

(12) 公開特許公報 (A)

(11) 特許出願公開番号

特開2001-142650

(P2001-142650A)

(43) 公開日 平成13年5月25日 (2001.5.25)

(51) Int.Cl.	識別記号	F I	特許出願公開番号
G 0 6 F 3/06	3 0 1	G 0 6 F 3/06	3 0 1 R 5 B 0 6 5
	3 0 5		3 0 1 S 5 D 0 6 6
	5 4 0		3 0 5 K
G 1 1 B 19/02	5 0 1	G 1 1 B 19/02	5 4 0
			5 0 1 F

審査請求 有 請求項の数 5 O L (全 7 頁) 最終頁に続く

(21) 出願番号 特願平11-327642

(22) 出願日 平成11年11月18日 (1999. 11. 18)

(71) 出願人 000004237

日本電気株式会社

東京都港区芝五丁目7番1号

(72) 発明者 浅野 聡宏

東京都港区芝五丁目7番1号 日本電気株式会社内

(74) 代理人 100079164

弁理士 高橋 勇

Fターム (参考) 5B065 BA01 CA13 CA30 CC08 CE12

EA04 EA25

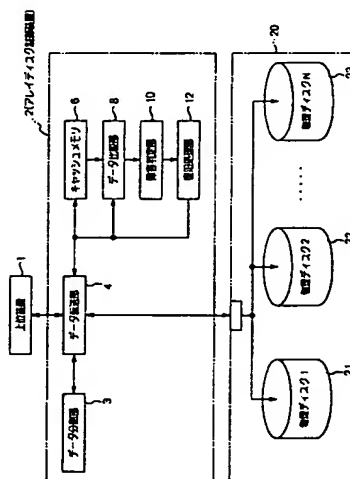
5D066 BA02 BA07 BA08

(54) 【発明の名称】 アレイディスク制御方法及び装置

(57) 【要約】

【課題】 不正データの発生及びディスク障害を早期に発見して長期間読み出されないデータであっても確実に保存すること。

【解決手段】 キャッシュメモリ6と、上位装置1から送信されたデータを物理ディスクの台数に応じて分割すると共に当該分割した分割データのバリティデータを生成するデータ分割部3と、この分割データ及びバリティデータを各物理ディスクにそれぞれ格納するデータ転送部4と、分割データ及びバリティデータとを各物理ディスクに格納した後、当該各物理ディスクから当該データを読み出してキャッシュメモリに格納された対応するデータと比較するデータ比較部8と、このデータ比較部8で一致しない部分がある場合には当該一致しない部分を記憶した物理ディスクに障害が発生したと判定する障害判定部10とを備えた。



(2) 001-142650 (P2001-50)

## 【特許請求の範囲】

【請求項1】 論理ディスクを構成する複数台の物理ディスクと、上位装置から送信されたデータを一時的に記憶するキャッシュメモリと、前記上位装置から送信されたデータを前記物理ディスクの台数に応じて分割すると共に当該分割した分割データのバリティデータを生成するデータ分割部と、このデータ分割部によって生成された分割データ及びバリティデータを前記各物理ディスクにそれぞれ格納するデータ転送部とを備えたディスクアレイ装置を使用して上位装置から送信されたデータを格納するディスクアレイ制御方法であって、

前記分割データ及びバリティデータの前記各物理ディスクへの格納処理に前後して前記上位装置に格納完了報告コマンドを送信する格納完了報告工程と、この格納完了報告工程の後前記キャッシュメモリから当該格納完了報告を行ったデータが削除されるまでの間に、前記各物理ディスクから当該データを読み出して前記キャッシュメモリに格納された対応するデータと比較するデータ比較工程と、このデータ比較工程で一致しない部分がある場合には当該一致しない部分を記憶した物理ディスクに障害が発生したと判定する障害判定工程とを備えたことを特徴とするアレイディスク制御方法。

【請求項2】 前記障害判定工程に続いて、当該障害と判定された物理ディスクに前記ディスクキャッシュに格納されたデータを再書き込みする再書き込み工程を備えたことを特徴とする請求項1記載のアレイディスク制御方法。

【請求項3】 前記障害判定工程に続いて、当該障害と判定された物理ディスクを縮退させて当該物理ディスクの使用を不可とする縮退工程を備えたことを特徴とする請求項1又は2記載のアレイディスク制御方法。

【請求項4】 前記格納完了報告工程に続いて、前記キャッシュメモリに格納されたデータのうち上位装置からのアクセス頻度の低いデータを選択する削除対象データ選択工程と、この削除対象データ選択工程で選択されたデータについて前記データ比較工程を実行させると共に前記障害判定工程によって障害がないと判定された場合に当該データを前記キャッシュメモリから削除する削除前比較工程とを備えたことを特徴とする請求項1記載のアレイディスク制御方法。

【請求項5】 上位装置から送信されたデータを一時的に記憶するキャッシュメモリと、前記上位装置から送信されたデータを物理ディスクの台数に応じて分割すると共に当該分割した分割データのバリティデータを生成するデータ分割部と、このデータ分割部によって生成された分割データ及びバリティデータを前記各物理ディスクにそれぞれ格納するデータ転送部とを備えたと共に、前記分割データ及びバリティデータとを各物理ディスクに格納した後、当該前記各物理ディスクから当該データを読み出して前記キャッシュメモリに格納された対応す

るデータと比較するデータ比較部と、このデータ比較工程で一致しない部分がある場合には当該一致しない部分を記憶した物理ディスクに障害が発生したと判定する障害判定部とを備えたことを特徴とするアレイディスク制御装置。

## 【発明の詳細な説明】

【0001】

【発明の属する技術分野】本発明は、アレイディスク制御方法及び装置に係り、特に、データに冗長性を持たせて複数の物理ディスクにデータを分割して格納するアレイディスク制御方法及び装置に関する。

【0002】

【従来の技術】従来より、上位装置から転送されたデータを分割し、さらに分割データのバリティデータを生成して、これらを別々の物理ディスクに格納することで、1つの物理ディスクに障害が発生した場合であってもデータを復旧可能な手法が用いられている。このようなデータ分割やバリティデータの生成については、例えばRAID3、5等にまとめられている。

【0003】しかしながら、上記従来例では、物理ディスクに障害が発生し、データの読み出しはできるが、データの内容が不正である場合には、どの物理ディスクに障害が発生したのかを判定することができない。これに対し、特開平9-305328号公報記載の手法では、分割データからバリティデータを生成するのみならず、巡回冗長検査情報(CRC)を付加している。そして、再生したデータのバリティが不正であった場合には、1台を障害ディスクとした場合の複数の組み合わせでデータを復元し、複数の復元データの内、CRCに基づいて正常であると判定された復元データを再生データとして上位装置に転送する。

【0004】

【発明が解決しようとする課題】しかしながら、上記従来例では、データの読み出し時に不正データのチェックを行うため、データ書き込み後長時間データを読み取らなかった場合に、不正データが発生した原因の解明が困難で特定できず、すると、データ書き込み後長時間経過した場合には、同時に複数台の物理ディスクで障害が発生してしまうことも想定され、すると、データの復旧が困難となる可能性が高い、という不都合があった。

【0005】

【発明の目的】本発明は、係る従来例の有する不都合を改善し、特に、不正データの発生及びディスク障害を早期に見出して長期間読み出されないデータであっても確実に保存することのできるアレイディスク制御方法及び装置を提供することを、その目的とする。

【0006】

【課題を解決するための手段】そこで、本発明では、論理ディスクを構成する複数台の物理ディスクと、上位装置から送信されたデータを一時的に記憶するキャッシュ

(3) 001-142650 (P2001- 記載)

メモリと、前記上位装置から送信されたデータを前記物理ディスクの台数に応じて分割すると共に当該分割した分割データのバリティデータを生成するデータ分割部と、このデータ分割部によって生成された分割データ及びバリティデータを前記各物理ディスクにそれぞれ格納するデータ転送部とを備えたディスクアレイ装置を使用して上位装置から送信されたデータを格納するディスクアレイ制御方法であって、分割データ及びバリティデータの前記各物理ディスクへの格納処理に前後して前記上位装置に格納完了報告コマンドを送信する格納完了報告工程と、この格納完了報告工程の後前記キャッシュメモリから当該格納完了報告を行ったデータが削除されるまでの間に、前記各物理ディスクから当該データを読み出して前記キャッシュメモリに格納された対応するデータと比較するデータ比較工程と、このデータ比較工程で一致しない部分がある場合には当該一致しない部分を記憶した物理ディスクに障害が発生したと判定する障害判定工程とを備えた、という構成を採っている。これにより前述した目的を達成しようとするものである。

【0007】本発明では、上位装置から送信されたデータをまずキャッシュメモリに格納する。キャッシュメモリに当該データを格納すると、上位装置に格納完了報告コマンドを送信する。そして、キャッシュメモリに格納されたデータを分割すると共に、バリティデータを生成し、それぞれ複数の物理ディスクに格納する。上位装置への格納完了報告は、この分割データ及びバリティデータの物理ディスクへの格納が完了した時に行うようにしてもよい。

【0008】続いて、上位装置から他のI/O要求がある場合にはそのI/O要求を処理する。そして、上記各物理ディスクに格納したデータと同一のデータをキャッシュメモリから削除する前までに、当該物理ディスクへのデータの格納が正常に完了したか否かの判定を行う。すなわち、データ比較工程では、各物理ディスクから当該データを読み出して前記キャッシュメモリに格納された対応するデータと比較する。そして、障害判定工程では、データ比較工程で一致しない部分がある場合には当該一致しない部分を記憶した物理ディスクに障害が発生したと判定する。このように、キャッシュメモリに格納されたデータと各物理ディスクに格納されたデータとを直接比較することにより、各物理ディスク毎に不正データの発生の有無を確認する。

【0009】

【発明の実施の形態】以下、本発明の実施の形態を図面を参照して説明する。

【0010】図1は本発明によるアレイディスク制御装置の実施形態の構成を示すブロック図である。図1に示すように、アレイディスク制御装置2は、上位装置1から送信されたデータを一時的に記憶するキャッシュメモリ6と、前記上位装置1から送信されたデータを前記物

理ディスクの台数に応じて分割すると共に当該分割した分割データのバリティデータを生成するデータ分割部3と、このデータ分割部によって生成された分割データ及びバリティデータを前記各物理ディスクにそれぞれ格納するデータ転送部4とを備えている。データ分割部3は、データ転送部3に併設せず、データ転送部4の一機能としてもよい。

【0011】アレイディスク制御装置2はさらに、分割データ及びバリティデータとを各物理ディスクに格納した後、当該前記各物理ディスクから当該データを読み出して前記キャッシュメモリに格納された対応するデータと比較するデータ比較部8と、このデータ比較部8で一致しない部分がある場合には当該一致しない部分を記憶した物理ディスクに障害が発生したと判定する障害判定部10とを備えている。

【0012】アレイディスク制御装置は、データ転送部4を介して、上位装置1と論理ディスク20を構成する各物理ディスク21、22、23と接続されている。ここでは、物理ディスクは、物理ディスク1から物理ディスクNまでのN個設けられている。

【0013】ディスク転送部4は、上位装置1とのデータの転送を行い、キャッシュメモリ6へのデータの読み取り/書き込み及び論理ディスク20へのデータの転送を行う。キャッシュメモリは、物理ディスクよりアクセスが速い例えばRAM等の記憶媒体を使用する。

【0014】データ比較部8は、上位装置から受け取った書き込みデータがキャッシュメモリ6に格納されている状態で、各物理ディスクに格納したデータを読み出してキャッシュメモリ6に格納されているデータと比較する。このとき、不一致部分があれば、正常に書き込まれていないこととなる。そして、正常に書き込まれなかった部分のデータが、どの物理ディスクに書き込まれたデータであるかは、データ分割部3による分割ルールに基づいて判明する。このため、比較したデータに不一致部分があれば、不一致を生じさせた不正データがどの物理ディスクに格納されていたかが判明する。

【0015】従って、障害判定部10は、データ比較部8で一致しない部分がある場合には当該一致しない部分を記憶した物理ディスクに障害が発生したと判定する。これにより、データの読み出しは可能ではあるが、正常に書き込みが行われないような障害が発生した物理ディスクを特定し、復旧処理を行うことができる。すると、長年に渡って読み出されないデータであっても、データの格納時に正確に格納されたか否かを確認するため、例えば複数台の物理ディスクが同時に障害を有しデータの復旧ができなくなる事態を防止することができる。また、データの格納時に正常に書き込みができなかった物理ディスクが特定されるため、当該物理ディスクの装置番号等を上位装置に書き込み障害情報として通知することで、不正データが生じた環境を維持したまま不正デー

(4) 001-142650 (P2001-eJ50)

タが生じた原因を説明することができ、従って、長期に渡った後に不正データが生じた場合と比較して、障害の原因説明が容易となる。

【0016】復旧処理部12は、障害判定部10によって障害が発生した物理ディスクが特定されると、例えば、キャッシュメモリ6から不正データ部分を含む分割データを読み出して、再書き込み処理を1回又は数回繰り返して、再度キャッシュメモリに格納したデータと比較することで不正データが生じたか否かを判定する。このとき、当該不正データを生じさせた物理ディスクの異なるセクタへ書き込みを行うようにしても良い。このような再書き込みを行っても不正データが生じる場合には、当該不正データを生じさせた物理ディスクを縮退させる。縮退処理は、他のデータ全ての復元を行った後、当該物理ディスクを切り離す等の処理となる。

【0017】図2は、上位装置（ホストプロセッサ）から書き込みデータとして送信されるデータの分割の例を示す説明図である。キャッシュメモリ6には、図2

(A)に示すようなホストプロセッサから受け取ったデータをそのまま格納する。一方論理ディスクに記録するデータは、ホストプロセッサから受け取ったデータをデータ分割部3で分割して各物理ディスク1（符号21）、物理ディスク2（符号22）、物理ディスクN（符号23）に記録する。N台の物理ディスクで構成される論理ディスク20にデータを記録する場合、図2(B)に示すように、データをN-1個に分割して分割データ1（符号31）、分割データ2（符号32）から分割データN-2（符号33）、分割データN-1（符号34）を生成する。そして、分割したデータのバリティデータ35を生成する。図2(C)に示すように、各物理ディスク1乃至Nには分割されたデータ1乃至N-1と、生成したバリティデータ35が書き込まれる。データを読みとる場合、分割したデータ1乃至N-1とバリティデータ35を全て読みとり、バリティチェックを行ってデータの妥当性を確認し、バリティデータ35を除いたものをホストプロセッサ1に転送する。

【0018】次に、図3を参照して図1に示したアレイドスク制御装置を使用してアレイドスクを制御する動作例を説明する。図3は本実施形態の動作例を示すフローチャートである。まず、データ転送部4が、上位装置（ホストプロセッサ）1から書き込みデータを受け取り（ステップS1）、キャッシュメモリ6へ当該データを格納する（ステップS2）。さらに、データ分割部3は、当該データを分割してバリティデータ35を生成する（ステップS3）。データ転送部4は、この分割データ31乃至34及びバリティデータ35を各物理ディスク21乃至23へ格納する（ステップS4）。続いて、データ転送部4は、ホストプロセッサ1へ格納完了報告を行う（ステップS5、格納完了報告工程）。具体的には、格納完了報告を示すコマンドを上位装置の入出力イ

ンタフェースへ送信する。この上位装置への格納完了報告は、実際に各物理ディスクにデータを格納した直後ではなく、ステップS2のキャッシュメモリ6への格納が完了した直後に行うようにしても良い。

【0019】図3に示す例では、格納完了報告S5に続いて、すなわち、上位装置を切り離した後、物理ディスクに格納したデータとキャッシュメモリ6に格納されているデータとの比較を行う（データ比較工程S6乃至S8）。データ転送部4は、物理ディスクへ格納したデータを読み出す（ステップS7）。一方、データ比較部8は、キャッシュメモリ6から同一データであったデータを読み出し、バリティデータを生成する（ステップS8）。そして、データ比較部8は、この物理ディスクから読み出したデータとキャッシュメモリ6に格納されたデータとを比較する（ステップS9）。図3に示す例では、データ比較工程が、キャッシュメモリから読み出したデータのバリティデータを生成する工程を備えている。これにより、バリティデータを格納したデータの不正の有無を判定することができる。

【0020】両データの比較の結果、一致していれば書き込みは正常に行われているため、データの書き込み処理を終了し、一方、両データに相違点がある場合には、当該一致しない部分を記憶した物理ディスクに障害が発生したと判定する（ステップS10、障害判定工程）。そして、一方、両データに相違点がある場合には、復旧処理部12を起動して復旧処理を行う（ステップS11、復旧処理工程）。

【0021】復旧処理工程S11は、当該障害と判定された物理ディスクに前記ディスクキャッシュに格納されたデータを再書き込みする再書き込み工程や、当該障害と判定された物理ディスクを縮退させて当該物理ディスクの使用を不可とする縮退工程を備えると良い。

【0022】図3に示す例では、上位装置1への格納完了報告S5が完了した直後に、物理ディスクへのデータの格納の直後にデータ比較工程を位置づけている。上位装置1からのデータの書き込み要求及び再生要求（I/O要求）が連続的であれば、図3に示すようにデータの書き込みの直後にデータ比較工程を実施することが好ましい。一方、上位装置のI/O要求が間欠的であれば、I/O要求を優先実行し、I/O要求の待機時間内を活用してデータ比較工程を実施するようにしても良い。この場合、本実施例ではキャッシュメモリ6に格納されたデータを利用するため、このキャッシュメモリ6に格納したデータを削除する前までには、データの比較を完了させなければならない。

【0023】図4は、キャッシュメモリ6からデータを削除する直前までにデータ比較を行う処理例を示すフローチャートである。図4に示す例では、格納完了報告工程S5に続いて、前記キャッシュメモリに格納されたデータのうち上位装置からのアクセス頻度の低いデータを

(5) 001-142650 (P2001-0050)

選択する削除対象データ選択工程S24と、この削除対象データ選択工程S24で選択されたデータについて前記データ比較工程S9を実行させると共に前記障害判定工程S10によって障害がないと判定された場合に当該データを前記キャッシュメモリから削除する削除前比較工程S27とを備えている。

【0024】具体的には、まず、上位装置からのI/O要求の有無を確認する(ステップS21)。I/O要求が書き込みである場合、キャッシュメモリ6に十分な容量が無ければ実行できなくなるため、キャッシュの容量が十分であるか否かを確認する(ステップS22)。キャッシュ容量が十分であれば、I/O処理を実行する(ステップS23)。すなわち、データの書き込みが要求されれば、図3に示すステップS1乃至S5を実行し、データの読み出しを要求された場合には、キャッシュメモリ6に該当するデータがあれば当該データを上位装置に送信し、一方キャッシュメモリ6に該当するデータが無ければ物理ディスクからデータを読み出して復元の上、上位装置に転送する。データ読み出し要求時に、キャッシュメモリ6に当該データが格納されていた場合には、当該データのアクセス頻度を示す値を増加させるようにすると良い。

【0025】上位装置からのI/O要求を待機している場合や、キャッシュ容量が不十分である場合には、まず、キャッシュメモリ6に格納されたデータの内、アクセス頻度の低いデータを検索する(ステップS24)。このデータの検索・特定は、削除すべきデータの特定であるため、アクセス頻度の他、一定時間経過したデータを選択するなど、アレイディスクの分野で一般的な他の手法を採用することができる。

【0026】削除候補となるデータが特定されると、当該データについて各ディスクに格納されたデータと比較する(ステップS25)。すなわち、図3に示すステップS6乃至S9を実行する。そして、両データが一致していれば、当該データは正常に格納されているため、当該キャッシュメモリに格納されたキャッシュデータを削除する(ステップS27)。キャッシュメモリに十分な空き容量がある場合には、削除可能であることを当該キャッシュメモリを管理する図示しないキャッシュメモリ管理部へ通知するようにしても良い。

【0027】一方、データが一致しない場合には、図3に復旧処理S11を実行する。

【0028】上述したように本実施形態によると、キャッシュ上のデータとディスクに書き込まれたデータを比較することにより、上位装置からの書込データが複数の物理ディスクすべてに正しく書き込まれたか否かを判定

することができ、すると、分割したデータの書き付けを未然に防ぐ事ができるようになる。

【0029】

【発明の効果】本発明は以上のように構成され機能するので、これによると、障害判定工程が、データ比較工程で一致しない部分がある場合には当該一致しない部分を記憶した物理ディスクに障害が発生したと判定するため、各物理ディスク毎に不正データの発生の有無をデータの書き込み直後に確認することができ、すると、物理ディスクの障害発生原因の解明が容易で且つ当該データを長期に渡って確実に保存することができ、しかも、キャッシュメモリに格納されたデータと物理ディスクに格納されたデータとを直接比較することで不正データを特定するため、不正データが格納されていた物理ディスクなどの物理ディスクであるかを即座に特定することができ、このため、障害復旧処理の選択肢が増え、例えば、複数回再書き込みを行った後にデータの不正が解消しない場合に当該物理ディスクについて縮退処理を行うなど、物理ディスクの障害の早期発見及び早期復旧を図ることができる、という従来にない優れたアレイディスク制御方法を提供することができる。

【図面の簡単な説明】

【図1】本発明の第一の実施形態の構成を示すブロック図である。

【図2】図1に示した各構成要素が扱うデータの構造を示す説明図であり、図2(A)は上位装置から転送されるデータの構造を示す図で、図2(B)はディスク分割部によって生成されるデータの構造を示す図で、図2(C)は各物理ディスクに格納されるデータの構造を示す図である。

【図3】図1に示した構成でのデータ書き込み処理の一例を示すフローチャートである。

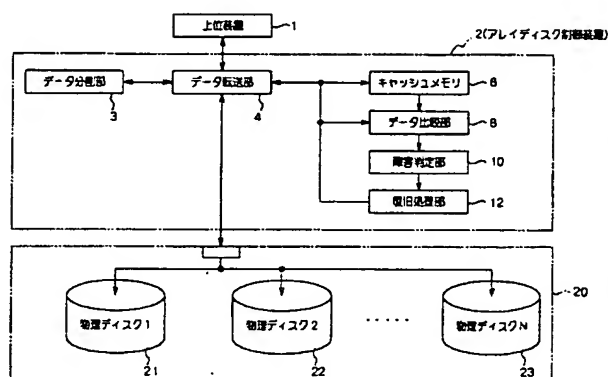
【図4】図1に示した構成での他のデータ書き込み処理の例を示すフローチャートである。

【符号の説明】

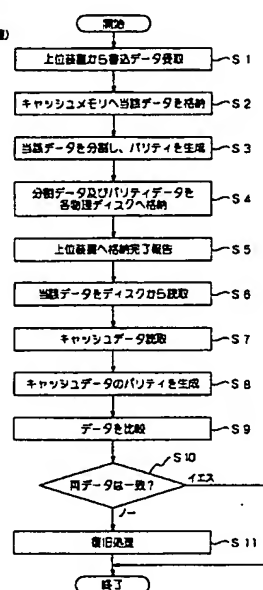
- 1 上位装置(ホストプロセッサ)
- 2 アレイディスク制御装置
- 3 データ分割部
- 4 データ転送部
- 6 キャッシュメモリ
- 8 データ比較部
- 10 障害判定部
- 12 復旧処理部
- 20 論理ディスク
- 21, 22, 23 論理ディスクを構成する物理ディスク

(6) 001-142650 (P2001-8) 簡略

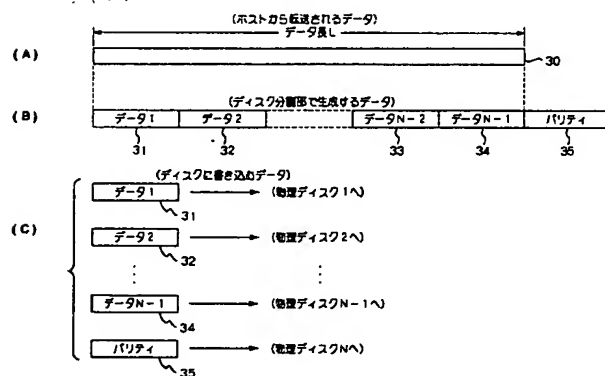
【図1】



【図3】

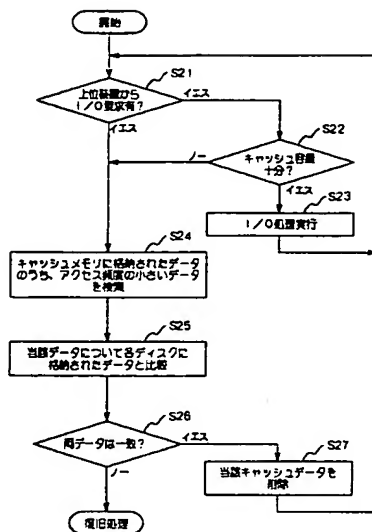


【図2】



(7) 001-142650 (P2001-0750)

【図4】



フロントページの続き

(51) Int. Cl.<sup>7</sup>

G11B 20/18

識別記号

501

522

570

572

FI

G11B 20/18

ターミナル (参考)

501B

522Z

570Z

572B

572F